

# PERSISTENT HOMOLOGY FOR METRIC MEASURE SPACES, AND ROBUST STATISTICS FOR HYPOTHESIS TESTING AND CONFIDENCE INTERVALS

ANDREW J. BLUMBERG, ITAMAR GAL, MICHAEL A. MANDELL,  
AND MATTHEW PANCIA

**ABSTRACT.** We study distributions of persistent homology barcodes associated to taking subsamples of a fixed size from metric measure spaces. We show that such distributions provide robust invariants of metric measure spaces, and illustrate their use in hypothesis testing and providing confidence intervals for topological data analysis.

## 1. INTRODUCTION

Topological data analysis assigns homological invariants to data presented as a finite metric space (a “point cloud”). If we imagine this data as measurements sampled from some abstract universal space  $M$ , the structure of that space is a metric measure space, having a notion both of distance between points and a notion of probability for the sampling. The usual homological approach to samples is to assign a simplicial complex and compute its homology. The construction of the associated simplicial complex for a point cloud depends on a choice of scale parameter. The insight of “persistence” is that one should study homological invariants that encode change across scales; the correct scale parameter is a priori unknown. As such, a first approach to studying the homology of  $M$  from the samples is to simply compute the persistent homology of the sampled point cloud.

We can gain some perspective from imagining that we could make measurements on  $M$  directly and interpret these measurements in terms of random sample points. With this in mind, we immediately notice some defects with homology and persistent homology as invariants of  $M$ . While the homology of  $M$  captures information about the global topology of the metric space, the probability space structure plays no role. This has bearing even if we assume  $M$  is a compact Riemannian manifold and the probability measure is the volume measure for the metric: handles which are small represent subsets of low probability but contribute to the homology in the same way as large handles. In this particular kind of example, persistent homology can identify this type of phenomenon (by encoding the scales at which homological features exist); however, in a practical context, the metric on the sample may be ad hoc (e.g., [3]) and less closely related to the probability measure. In this case, we could have handles that are medium size with respect to the metric but still low probability with respect to the measure. Homology and persistent homology have no mechanism for distinguishing low probability features from high probability features. A closely related issue is the effect of small amounts of noise (e.g., a

---

*Date:* March 8, 2013.

The authors were supported in part by DARPA YFA award N66001-10-1-4043.

situation in which a fraction of the samples are corrupted). A small proportion of bad samples can arbitrarily change the persistent homology. These two kinds of phenomena are linked, insofar as decisions about whether low probability features are noise or not is part of data analysis.

The disconnect with the underlying probability measure presents a significant problem when trying to adapt persistent homology to the setting of *hypothesis testing* and *confidence intervals*. Hypothesis testing involves making quantitative statements about the probability that the persistent homology computed from a sampling from a metric measure space is consistent with (or refutes) a hypothesis about the actual persistent homology. Confidence intervals provide a language to understand the variability in estimates introduced by the process of sampling. Because low probability features and a small proportion of bad samples can have a large effect on persistent homology computations, the persistent homology groups make poor test statistics for hypothesis testing and confidence intervals. To obtain useable test statistics, we need to develop invariants that better reflect the underlying measure and are less sensitive to large perturbation. To be precise about this, we use the statistical notion of robustness.

A statistical estimator is *robust* when its value cannot be arbitrarily perturbed by a constant proportion of bad samples. For instance, the sample mean is not robust, as a single extremely large sample value can dominate the result. On the other hand, the sample median is robust. As we discuss in Section 4, persistent homology is not robust. A small number of bad samples can cause large changes in the persistent homology, essentially as a reflection of the phenomenon of large metric low probability handles (including spurious ones).

Using the idea of an underlying metric measure space  $M$ , formally the process of sampling amounts to considering random variables on the probability space  $M^n = M \times \cdots \times M$  equipped with the product probability measure. The  $k$ -th persistent homology of a size  $n$  sample is a random variable on  $M^n$  taking values in the set  $\mathcal{B}$  of finite *barcodes* [23], where a barcode is essentially a multiset of intervals of the form  $[a, b)$ . The set  $\mathcal{B}$  of barcodes is equipped with a metric, the bottleneck metric [7], and we show in Section 3 that it is separable and that its completion  $\overline{\mathcal{B}}$  is also a space of barcodes. Then  $\overline{\mathcal{B}}$  is Polish, i.e., complete and separable, which makes it amenable to probability theory (see also [18] for such results). In particular, various metrics on the set of distributions on  $\overline{\mathcal{B}}$  metrize weak convergence, including the Prohorov metric  $d_{Pr}$  and the Wasserstein metric  $d_W$ . We consider the following probability distribution on barcode space  $\overline{\mathcal{B}}$ :

**Definition 1.1.** For a metric measure space  $(X, \partial_X, \mu_X)$  and fixed  $n, k \in \mathbb{N}$ , define the  $k$ th  $n$ -sample persistent homology as

$$\Phi_k^n(X, \partial_X, \mu_X) = (\text{PH}_k)_*(\mu_X^{\otimes n}),$$

the probability distribution on the set of barcodes  $\overline{\mathcal{B}}$  induced by pushforward along  $\text{PH}_k$  from the product measure  $\mu_X^n$  on  $X^n$ .

In other words,  $\Phi_k^n$  is the probability measure on the space of barcodes where the probability of a subset  $A$  is the probability that a size  $n$  sample from  $M$  has  $k$ -th persistent homology landing in  $A$ .

Although complicated,  $\Phi_k^n(M)$  is a continuous invariant of  $M$  in the following sense. The moduli space of metric measure spaces admits a metric (in fact several) that combine the ideas of the Gromov-Hausdorff distance on compact metric

spaces and weak convergence of probability measures [22]. We follow [11], and use the *Gromov-Prohorov metric*,  $d_{GPr}$ . We prove the following theorem in Section 5 (where it is restated as Theorem 5.2).

**Theorem 1.2.** *Let  $(X, \partial_X, \mu_X)$  and  $(Y, \partial_Y, \mu_Y)$  be compact metric measure spaces. Then we have the following inequality relating the Prohorov and Gromov-Prohorov metrics:*

$$d_{Pr}(\Phi_k^n(X, \partial_X, \mu_X), \Phi_k^n(Y, \partial_Y, \mu_Y)) \leq n d_{GPr}((X, \partial_X, \mu_X), (Y, \partial_Y, \mu_Y)).$$

As a consequence of the continuity implied by the previous theorem, we can use  $\Phi_k^n$  to develop robust statistics: If we change  $M$  by adjusting the metric arbitrarily on  $\epsilon$  probability mass to produce  $M'$ , then the Gromov-Prohorov distance satisfies  $d_{GPr}(M, M') \leq \epsilon$ .

A first question that arises is how to interpret  $\Phi_k^n$  in practice, where we are given a large finite sample  $S$  which we regard as drawn from  $M$ . Making  $S$  a metric measure space via the subspace metric from  $M$  and the empirical measure, we can compute  $\Phi_k^n(S)$  as an approximation to  $\Phi_k^n(M)$ . This procedure is justified by the fact that as the sample size increases, the empirical metric converges almost surely in  $d_{GPr}$  to  $M$ ; see Lemma 5.4. (This kind of approximation is intimately connected with resampling methodology, a topic we study in the paper [1].)

A problem with  $\Phi_k^n$  is that it can be hard to interpret or summarize the information contained in a distribution of barcodes, unlike distributions of numbers for which there are various moments (e.g., the mean and the variance) which provide concise summaries of the distribution. One approach is to develop “topological summarizations” of distributions of barcodes, a subject we pursue in the paper [2]. In this paper, we instead consider cruder invariants which take values in  $\mathbb{R}$ . One such invariant is the distance with respect to a reference distribution on barcodes  $\mathcal{P}$ , chosen to represent a hypothesis about the persistent homology of  $M$ .

**Definition 1.3.** Let  $(X, \partial_X, \mu_X)$  be a compact metric measure space and let  $\mathcal{P}$  be a fixed reference distribution on  $\bar{\mathcal{B}}$ . Fix  $k, n \in \mathbb{N}$ . Define the homological distance on  $X$  relative to  $\mathcal{P}$  to be

$$\text{HD}_k^n((X, \partial_X, \mu_X), \mathcal{P}) = d_{Pr}(\Phi_k^n(X, \partial_X, \mu_X), \mathcal{P}).$$

We also produce a robust statistic  $\text{MHD}_k^n$  related to  $\text{HD}_k^n$  without first computing the distribution  $\Phi_k^n$ . To construct  $\text{MHD}_k^n$ , we start with a reference bar code and compute the median distance to the barcodes of subsamples.

**Definition 1.4.** Let  $(X, \partial_X, \mu_X)$  be a compact metric measure space and let  $\mathcal{P}$  be a fixed reference barcode  $B \in \bar{\mathcal{B}}$ . Fix  $k, m \in \mathbb{N}$ . Let  $\mathcal{D}$  denote the distribution on  $\mathbb{R}$  induced by applying  $d_B(B, -)$  to the barcode distribution  $\Phi_k^n(X, \partial_X, \mu_X)$ . Define the median homological distance relative to  $\mathcal{P}$  to be

$$\text{MHD}_k^n((X, \partial_X, \mu_X), \mathcal{P}) = \text{median}(\mathcal{D}).$$

The use of the median rather than the mean in the following definition ensures that we compute a robust statistic. To be precise, for functions from finite metric spaces to metric spaces, we use the following definition of robustness.

**Definition 1.5.** Let  $f$  be a function from finite metric spaces to a metric space  $(B, d)$ . We say that  $f$  is robust with robustness coefficient  $r > 0$  if for any non-empty finite metric space  $(X, \partial)$ , there exists a bound  $\delta$  such that for any isometry of  $X$

into a finite metric space  $(X', \partial')$ ,  $|X'|/|X| < 1+r$  implies  $d(f(X, \partial), f(X', \partial')) < \delta$ , where  $|X|$  denotes the number of elements of  $X$ .

For example, under the analogous definition on finite multi-subsets of  $\mathbb{R}$  (in place of finite metric spaces), median defines a function to  $\mathbb{R}$  that is robust with robustness coefficient  $1 - \epsilon$  for any  $\epsilon$  since expanding a multi-subset  $X$  to a larger one  $X'$  with fewer than twice as many elements will not change the median by more than the diameter of  $X$ . Similarly, for a finite metric space  $X$ , expanding  $X$  to  $X'$ , the proportion of  $n$ -element samples of  $X'$  which are samples of  $X$  is  $(|X|/|X'|)^n$ ; when this number is more than  $1/2$ , the median value of any function  $f$  on the set of  $n$ -element samples of  $X'$  is then bounded by the values of  $f$  on  $n$ -element samples of  $X$ . Since  $(N/(N + rN))^n > 1/2$  for  $r < 2^{1/n} - 1$ , any such function  $f$  will be robust with robustness coefficient  $r$  satisfying this bound, and in particular for  $r = (\ln 2)/n$ .

**Theorem 1.6.** *For any  $n, k, \mathcal{P}$ , the function  $\text{MHD}_k^n(-, \mathcal{P})$  from finite metric spaces (given the uniform probability measure) to  $\mathbb{R}$  is robust with robustness coefficient  $> (\ln 2)/n$ .*

The function  $\Phi_k^n$  from finite metric spaces to distributions on  $\bar{\mathcal{B}}$  and the function  $\text{HD}_k^n$  from finite metric spaces to  $\mathbb{R}$  are robust for any robustness coefficient for trivial reasons since the Gromov-Prohorov metric is bounded. However, for these functions we can give explicit uniform estimates for how much these functions change when expanding  $X$  to  $X'$  just based on  $|X'|/|X|$ . We introduce the following notion of uniform robustness strictly stronger than the notion of robustness.

**Definition 1.7.** Let  $f$  be a function from finite metric spaces to a metric space  $(B, d)$ . We say that  $f$  is uniformly robust with robustness coefficient  $r > 0$  and estimate bound  $\delta$  if for any non-empty finite metric space  $(X, \partial)$  and any isometry of  $(X, \partial)$  into a finite metric space  $(X', \partial')$ ,  $|X'|/|X| < 1 + r$  implies  $d(f(X, \partial), f(X', \partial')) < \delta$ .

Uniform robustness gives a uniform estimate on the change in the function from expanding the finite metric space. For example, the median function does not satisfy the analogous notion of uniform robustness for functions on finite multi-subsets of  $\mathbb{R}$ . We show in Section 5 that  $\Phi_k^n$  and  $\text{HD}_k^n$  satisfy this stronger notion of uniform robustness.

**Theorem 1.8.** *For fixed  $n, k$ ,  $\Phi_k^n$  is uniformly robust with robustness coefficient  $r$  and estimate bound  $nr/(1+r)$  for any  $r$ . For fixed  $n, k, \mathcal{P}$ ,  $\text{HD}_k^n(-, \mathcal{P})$  is uniformly robust with robustness coefficient  $r$  and estimate bound  $nr/(1+r)$  for any  $r$ .*

As with  $\Phi_k^n$  itself, the law of large numbers and the convergence of empirical metric measure spaces tells us that given a sufficiently large finite sample  $S \subset M$ , we can approximate  $\text{HD}_k^n$  and  $\text{MHD}_k^n$  of the metric measure space  $M$  in a robust fashion from the persistent homology computations from  $S$ . (See Lemmas 5.4, 6.5, and 6.8 below.)

In light of the results on robustness and asymptotic convergence, HD, MHD, and  $\Phi$  (as well as various distributional invariants associated to  $\Phi$ ) provide good test statistics for hypothesis testing. Furthermore, one of the benefits of the numerical statistics  $\text{HD}_k^n$  and  $\text{MHD}_k^n$  is that we can use standard techniques to obtain confidence intervals, which provide a means for understanding the reliability of analyses

of data sets. We discuss hypothesis testing and the construction of confidence intervals in Section 6, and explore examples in Sections 7 and 8. In this paper we primarily focus on analytic methods which use asymptotic normality results; however, these statistics are well-suited for the construction of resampling confidence intervals. In the follow-up paper [1] we establish the asymptotic consistency of the bootstrap for  $\text{HD}_k^n$  and  $\text{MHD}_k^n$ .

We regard this paper as a first step towards providing a foundation for the integration of standard statistical methodology into computational algebraic topology. Our goal is to provide tools for practical use in topological data analysis. In a subsequent paper [1], we provide theoretical support for the use of resampling methods to understand the distributional invariants  $\Phi_k^n$ ,  $\text{HD}_k^n$ , and  $\text{MHD}_k^n$ . We also study elsewhere topological summarizations in order to understand distributions of barcodes [2].

The paper is organized as follows. In Section 2, we provide a rapid review of the necessary background on simplicial complexes, persistent homology, and metric measure spaces. In Section 3, we study the space of barcodes, establishing foundations needed to work with distributions of barcodes. In Section 4, we discuss the robustness of persistent homology. In Section 5, we study the properties of  $\Phi_k^n$ ,  $\text{MHD}_k^n$ , and  $\text{HD}_k^n$  and prove Theorem 1.2. We discuss hypothesis testing and confidence intervals in Section 6, which we illustrate with synthetic examples in Section 7. Section 8 applies these ideas to the analysis of the natural images data in [3].

## 2. BACKGROUND

**2.1. Simplicial complexes associated to point clouds.** Computational algebraic topology proceeds by assigning a simplicial complex (which usually also depends on a scale parameter  $\epsilon$ ) to a finite metric space  $(X, \partial)$ . Recall that a simplicial complex is a combinatorial model of a topological space, defined as a collection of nonempty finite sets  $Z$  such that for any set  $Z \in \mathcal{Z}$ , every nonempty subset of  $Z$  is also in  $\mathcal{Z}$ . Associated to such a simplicial complex is the “geometric realization”, which is formed by gluing standard simplices of dimension  $|Z| - 1$  via the subset relations. (The standard  $n$ -simplex has  $n + 1$  vertexes.) The most basic and widely used construction of a simplicial complex associated to a point cloud is the Vietoris-Rips complex:

**Definition 2.1.** For  $\epsilon \in \mathbb{R}$ ,  $\epsilon \geq 0$ , the Vietoris-Rips complex  $\text{VR}_\epsilon(X)$  is the simplicial complex with vertex set  $X$  such that  $[v_0, v_1, \dots, v_n]$  is an  $n$ -simplex when for each pair  $v_i, v_j$ , the distance  $\partial(v_i, v_j) \leq \epsilon$ .

Observe that the Vietoris-Rips complex is determined by its 1-skeleton. The construction is functorial in the sense that for a continuous map  $f: X \rightarrow Y$  with Lipschitz constant  $\kappa$  and for  $\epsilon \leq \epsilon'$ , there is a commutative diagram

$$(2.2) \quad \begin{array}{ccc} \text{VR}_\epsilon(X) & \longrightarrow & \text{VR}_{\kappa\epsilon}(Y) \\ \downarrow & & \downarrow \\ \text{VR}_{\epsilon'}(X) & \longrightarrow & \text{VR}_{\kappa\epsilon'}(Y). \end{array}$$

The Vietoris-Rips complex is easy to compute, but can be unmanageably large; for a set of points  $Y = \{y_1, y_2, \dots, y_n\}$  such that  $\partial(y_i, y_j) \leq \epsilon$ , every subset of  $Y$

specifies a simplex of the Vietoris-Rips complex. More closely related to classical constructions in algebraic topology is the Čech complex.

**Definition 2.3.** For  $\epsilon \in \mathbb{R}$ ,  $\epsilon \geq 0$ , the Čech complex  $C_\epsilon(X)$  is the simplicial complex with vertex set  $X$  such that  $[v_0, v_1, \dots, v_n]$  is an  $n$ -simplex when the intersection

$$\bigcap_{0 \leq i \leq n} B_{\frac{\epsilon}{2}}(v_i)$$

is non-empty, where here  $B_r(x)$  denotes the  $r$ -ball around  $x$ .

The Čech complex has analogous functoriality properties to the Vietoris-Rips complex. In Euclidean space, the topological Čech complex associated to a cover of a paracompact topological space satisfies the nerve lemma: if the cover consists of contractible spaces such all finite intersections are contractible or empty, the resulting simplicial complex is homotopy equivalent to the original space.

*Remark 2.4.* It is also often very useful to define complexes with the vertices restricted to a small set of landmark points; the weak witness complex is perhaps the best example of such a simplicial complex [21]. We discuss this construction further in Section 8, as it is important in the applications.

The theory we develop in this paper is relatively insensitive to the specific details of the construction of a simplicial complex associated to a finite metric space (and scale parameter). For reasons that will become evident when we discuss persistence in Subsection 2.3 below, the only thing we require is a procedure for assigning a complex to  $((M, \partial), \epsilon)$  that is functorial in the vertical maps of diagram (2.2) for  $\kappa = 1$ .

**2.2. Homological invariants of point clouds.** In light of the previous subsection, given a finite metric space  $(X, \partial)$ , one defines the homology at the feature scale  $\epsilon$  to be the homology of a simplicial complex associated to  $(X, \partial)$ ; e.g.,  $H_*(\text{VR}_\epsilon(X))$  or  $H_*(C_\epsilon(X))$ . This latter definition is supported by the following essential consistency result, which is in line with the general philosophy that we are studying an underlying continuous geometric object via finite sets of samples.

**Theorem 2.5** (Niyogi-Smale-Weinberger [19]). *Let  $(M, \partial)$  be a compact Riemannian manifold equipped with an embedding  $\gamma: M \rightarrow \mathbb{R}^n$ , and let  $X \subset M$  be a finite sample drawn according to the volume measure on  $M$ . Then for any  $p \in (0, 1)$ , there are constants  $\delta$  (which depends on the curvature of  $M$  and the embedding  $\gamma$ ) and  $N_{\delta, p}$  such that if  $\epsilon < \delta$  and  $|X| > N_{\delta, p}$  then the probability that  $H_*(C_\epsilon(X)) \cong H_*(M)$  is an isomorphism is  $> p$ .*

In fact, Niyogi, Smale, and Weinberger prove an effective version of the previous result, in the sense that there are explicit numerical bounds dependent on  $p$  and a “condition number” which incorporates data about the curvature of  $M$  and the twisting of the embedding  $\gamma$ .

Work by Latschev provides an equivalent result for  $\text{VR}_\epsilon(X)$ , with somewhat worse bounds, defined in terms of the injectivity radius of  $M$  [16]. Alternatively, one can obtain consistency results for  $\text{VR}_\epsilon(X)$  using the fact that there are inclusions

$$C_\epsilon(X) \subseteq \text{VR}_\epsilon(X) \subseteq C_{2\epsilon}(X).$$

While reassuring, an unsatisfactory aspect of the preceding results is the dependence on a priori knowledge of the feature scale  $\epsilon$  and the details of the intrinsic

curvature of  $M$  and the nature of the embedding. A convenient way to handle the fact that it is often hard to know a good choice of  $\epsilon$  at the outset is to consider multi-scale homological invariants that encode the way homology changes as  $\epsilon$  varies. This leads us to the notion of persistent homology [10].

**2.3. Persistent homology.** Given a diagram of simplicial complexes indexed on  $\mathbb{N}$  (i.e., a direct system),

$$X_0 \rightarrow X_1 \rightarrow \dots \rightarrow X_n \rightarrow \dots,$$

there are natural inclusions  $H_*(X_i) \rightarrow H_*(X_j)$  for  $i \leq j$ .

We say that a class  $\alpha \in H_p(X_i)$  is *born* at time  $i$  if it is not in the image of  $H_k(X_j)$  for  $j < i$ , and we say a class  $\alpha \in H_k(X_i)$  *dies* at time  $i$  if the image of  $\alpha$  is 0 in  $H_k(X_j)$  for  $j \geq i$ . This information about the homology can be packaged up into an algebraic object:

**Definition 2.6.** Let  $\{X_i\}$  be a direct system of simplicial complexes. The  $p$ th persistent  $k$ th homology group of  $X_i$  is defined to be

$$H_{k,p}(X_i) = Z_k^i / (B_k^{i+p} \cap Z_k^i),$$

where  $Z$  and  $B$  denote the cycle and boundary groups respectively. Alternatively,  $H_{k,p}(X_i)$  is the image of the natural map

$$H_k(X_i) \rightarrow H_k(X_{i+p}).$$

When working over a field and in the presence of suitable finiteness hypotheses, barcodes provide a convenient reformulation of information from persistent homology. Specifically, assume that the direct system of simplicial complexes stabilizes at a finite stage and all homology groups are finitely-generated. Then a basic classification result of Zomorodian-Carlsson [23] describes the persistent homology in terms of a barcode, a multiset of non-empty intervals of the form  $[a, b) \subset \mathbb{R}$ . An interval in the barcode indicates the birth and death of a specific homological feature. For reasons we explain below, the barcodes appearing in our context will always have finite length intervals.

The Rips (or Čech) complexes associated to a point cloud  $(X, \partial_X)$  fit into this context by looking at a sequence of varying values of  $\epsilon$ :

$$\text{VR}_{\epsilon_1}(X) \rightarrow \text{VR}_{\epsilon_2}(X) \rightarrow \dots$$

We can do this in several ways, for example, using the fact that the Vietoris-Rips complex changes only at discrete points  $\{\epsilon_i\}$  and stabilizes for sufficiently large  $\epsilon$ , or just choosing and fixing a finite sequence  $\epsilon_i$  independently of  $X$ . The theory we present below makes sense for either of these choices, and we use the following notation.

**Notation 2.7.** Let  $(X, \partial_X)$  be a finite metric space. For  $k \in \mathbb{N}$ , denote the persistent homology of  $X$  as

$$\text{PH}_k((X, \partial_X)) = \text{PH}_{k,p}(\{\text{VR}_{\epsilon_{(-)}}(X)\})$$

for some chosen sequence  $0 < \epsilon_1 < \epsilon_2 < \dots$  and  $p \geq 0$ .

More generally, we can make analogous definitions for any functor

$$\Psi: \mathcal{M} \times \mathbb{R}_{>0} \rightarrow \text{sComp},$$

where  $\mathcal{M}$  is the category of finite metric spaces and metric maps and  $\text{sComp}$  denotes the category of simplicial complexes. We will call such a  $\Psi$  “good” when the homotopy type changes for only finitely values in  $\mathbb{R}$ . In this case, we can choose the directed system of values of  $\epsilon_i$  to contain these transition values.

We note that for large values of the parameter  $\epsilon$ ,  $VR_\epsilon(X)$  will be contractible. Therefore, if we use the reduced homology group in dimension 0, we get  $H_k(VR_\epsilon) = 0$  for all  $k$  for large  $\epsilon$ . The bar codes associated to these persistent homologies therefore have only finite length bars. For convenience in computation, we typically cut off  $\epsilon$  at a moderately high value before this breakdown occurs. The result is a truncation of the bar code to the cut-off point.

**2.4. Gromov-Hausdorff stability and the bottleneck metric.** By work of Gromov, the set of isometry classes of finite metric spaces admits a useful metric structure, the Gromov-Hausdorff metric. For a pair of finite metric spaces  $(X_1, \partial_1)$  and  $(X_2, \partial_2)$ , the Gromov-Hausdorff distance is defined as follows: For a compact metric space  $(Z, \partial)$  and closed subsets  $A, B \subset Z$ , the Hausdorff distance is defined to be

$$d_H^Z(A, B) = \max(\sup_{a \in A} \inf_{b \in B} \partial(a, b), \sup_{b \in B} \inf_{a \in A} \partial(a, b)).$$

One then defines the Gromov-Hausdorff distance between  $X_1$  and  $X_2$  to be

$$d_{GH}(X_1, X_2) = \inf_{Z, \gamma_1, \gamma_2} d_H^Z(X_1, X_2),$$

where here  $\gamma_1: X_1 \rightarrow Z$  and  $\gamma_2: X_2 \rightarrow Z$  are isometric embeddings.

Since the topological invariants we are studying ultimately arise from finite metric spaces, a natural question to consider is the degree to which point clouds that are close in the Gromov-Hausdorff metric have similar homological invariants. This question does not in general have a good answer in the setting of the homology of the point cloud, but in the context of persistent homology, Chazal, et al. [6, 3.1] provide a seminal theorem in this direction that we review as Theorem 2.9 below.

The statement of Theorem 2.9 involves a metric on the set of barcodes called the bottleneck distance and defined as follows. Recall that a barcode  $\{I_\alpha\}$  is a multiset of non-empty intervals. Given two non-empty intervals  $I_1 = [a_1, b_1]$  and  $I_2 = [a_2, b_2]$ , define the distance between them to be

$$d_\infty(I_1, I_2) = \|(a_1, b_1) - (a_2, b_2)\|_\infty = \max(|a_1 - a_2|, |b_1 - b_2|).$$

We also make the convention

$$d_\infty([a, b], \emptyset) = |b - a|/2$$

for  $b > a$  and  $d_\infty(\emptyset, \emptyset) = 0$ . For the purposes of the following definition, we define a matching between two barcodes  $B_1 = \{I_\alpha\}$  and  $B_2 = \{J_\beta\}$  to be a multi-subset  $C$  of the underlying set of

$$(B_1 \cup \{\emptyset\}) \times (B_2 \cup \{\emptyset\})$$

such that  $C$  does not contain  $(\emptyset, \emptyset)$  and each element  $I_\alpha$  of  $B_1$  occurs as the first coordinate of an element of  $C$  exactly the number of times (counted with multiplicity) of its multiplicity in  $B_1$ , and likewise for every element of  $B_2$ . We get a more intuitive but less convenient description of a matching using the decomposition of  $(B_1 \cup \{\emptyset\}) \times (B_2 \cup \{\emptyset\})$  into its evident four pieces: The basic data of  $C$  consists of multi-subsets  $A_1 \subset B_1$  and  $A_2 \subset B_2$  together with a bijection (properly accounting for multiplicities)  $\gamma: A_1 \rightarrow A_2$ ;  $C$  is then the (disjoint) union of the graph of  $\gamma$



viewed as a multi-subset of  $B_1 \times B_2$ , the multi-subset  $(B_1 - A_1) \times \{\emptyset\}$  of  $B_1 \times \{\emptyset\}$ , and the multi-subset  $\{\emptyset\} \times (B_2 - A_2)$  of  $\{\emptyset\} \times B_2$ . With this terminology, we can define the bottleneck distance.

**Definition 2.8.** The bottleneck distance between barcodes  $B_1 = \{I_\alpha\}$  and  $B_2 = \{J_\beta\}$  is

$$d_{\mathcal{B}}(B_1, B_2) = \inf_C \sup_{(I, J) \in C} d_\infty(I, J),$$

where  $C$  varies over all matchings between  $B_1$  and  $B_2$ .

Although expressed slightly differently, this agrees with the bottleneck metric as defined in [7, §3.1] and [6, §2.2]. On the set of barcodes  $\mathcal{B}$  with finitely many finite length intervals,  $d_{\mathcal{B}}$  is obviously a metric. More generally, for any  $p > 0$ , one can consider the  $\ell^p$  version of this metric,

$$d_{B,p}(B_1, B_2) = \inf_C \left( \sum_{(I, J) \in C} d_\infty(I, J)^p \right)^{1/p}.$$

For simplicity, we focus on  $d_{\mathcal{B}}$  in this paper, but analogues of our main theorems apply to these variant metrics as well.

We have the following essential stability theorem:

**Theorem 2.9** (Chazal, et. al. [6, 3.1]). *For each  $k$ , we have the bound*

$$d_{\mathcal{B}}(\text{PH}_k(X), \text{PH}_k(Y)) \leq d_{GH}(X, Y).$$

Note that truncating bar codes is a Lipschitz map  $\mathcal{B} \rightarrow \mathcal{B}$  with Lipschitz constant 1, so the bound above still holds when we use a large parameter cut-off in defining  $\text{PH}_k$ .

**2.5. Metric measure spaces and the Gromov-Prohorov distance.** To establish more robust convergence results, we work with suitable metrics on the set of compact metric measure spaces. Specifically, following [11, 17, 22] we use the idea of the Gromov-Hausdorff metric to extend certain standard metrics on distributions (on a fixed metric measure space) to a metric on the set of all compact metric measure spaces.

A basic metric of this kind is the Gromov-Prohorov metric [11]. (For the following formulas, see Section 5 of [11] and its references.) This is defined in terms of the standard Prohorov metric  $d_{Pr}$  (metrizing weak convergence of probability distributions) as

$$d_{GPr}((X, \partial_X, \mu_X), (Y, \partial_Y, \mu_Y)) = \inf_{(\phi_X, \phi_Y, Z)} d_{Pr}^{(Z, \partial_Z)}((\phi_X)_* \mu_X, (\phi_Y)_* \mu_Y),$$

where the inf is computed over all isometric embeddings  $\phi_X: X \rightarrow Z$  and  $\phi_Y: Y \rightarrow Z$  into a target metric space  $(Z, \partial_Z)$ .

It is very convenient to reformulate both the Gromov-Hausdorff and Gromov-Prohorov distances in terms of relations. For sets  $X$  and  $Y$ , a relation  $R \subset X \times Y$  is a correspondence if for each  $x \in X$  there exists at least one  $y \in Y$  such that  $(x, y) \in R$  and for each  $y' \in Y$  there exists at least one  $x' \in X$  such that  $(x', y') \in R$ . For a relation  $R$  on metric spaces  $(X, \partial_X)$  and  $(Y, \partial_Y)$ , we define the distortion as

$$\text{dis}(R) = \sup_{(x, y), (x', y') \in R} |\partial_X(x, x') - \partial_Y(y, y')|.$$

The Gromov-Hausdorff distance can be expressed as

$$d_{GH}((X, \partial_X), (Y, \partial_Y)) = \frac{1}{2} \inf_R \text{dis}(R),$$

where we are taking the infimum over all correspondences  $R \subset X \times Y$ .

Similarly, we can reformulate the Prohorov metric as follows. Given two measures  $\mu_1$  and  $\mu_2$  on a metric space  $X$ , let a coupling of  $\mu_1$  and  $\mu_2$  be a measure  $\psi$  on  $X \times X$  (with the product metric) such that  $\psi(X \times -) = \mu_2$  and  $\psi(- \times X) = \mu_1$ . Then we have

$$d_{Pr}(\mu_1, \mu_2) = \inf_{\psi} \inf \{ \epsilon > 0 \mid \psi \{ (x, x') \in X \times X \mid \partial_X(x, x') \geq \epsilon \} \leq \epsilon \}.$$

This characterization of the Prohorov metric turns out to be useful when working with the Gromov-Prohorov metric in light of the (trivial) observation that if  $d_{GPr}((X, \partial_X, \mu_X), (Y, \partial_Y, \mu_Y)) < \epsilon$  then there exists a metric space  $Z$  and embeddings  $\iota_1: X \rightarrow Z$  and  $\iota_2: Y \rightarrow Z$  such that  $d_{Pr}((\iota_1)_*\mu_X, (\iota_2)_*\mu_Y) < \epsilon$ .

### 3. PROBABILITY MEASURES ON THE SPACE OF BARCODES

This section introduces the spaces of barcodes  $\mathcal{B}_N$  and  $\bar{\mathcal{B}}$  used in the distributional invariants  $\Phi_k^n$  of Definition 1.1. These spaces are complete and separable under the bottleneck metric. This implies in particular that the Prohorov metric on the set of probability measures in  $\mathcal{B}_N$  or  $\bar{\mathcal{B}}$  metrizes convergence in probability, which justifies the perspective in the stability theorem 1.2 and the definition of the invariants  $\text{HD}_k^n(-, \mathcal{P})$  in Definition 1.3.

A barcode is by definition a multi-set of intervals, in our case of the form  $[a, b)$  for  $0 \leq a < b < \infty$ . The set  $\mathcal{I}$  of all intervals of this form is of course in bijective correspondence with a subset of  $\mathbb{R}^2$ . A multi-set  $A$  of intervals is a multi-subset of  $\mathcal{I}$ , which concretely is a function from  $\mathcal{I}$  to the natural numbers  $\mathbb{N} = \{0, 1, 2, 3, \dots\}$  which counts the number of multiples of each interval in  $A$ . We denote by  $|A|$  the cardinality of  $A$ , which we define as the sum of the values of the function  $\mathcal{I} \rightarrow \mathbb{N}$  specified by  $A$  (if finite, or countably or uncountably infinite, if not). The space  $\mathcal{B}$  of barcodes of the introduction is the set of multi-sets of intervals  $A$  such that  $|A| < \infty$ . We have the following important subsets of  $\mathcal{B}$

**Definition 3.1.** For  $N \geq 0$ , let  $\mathcal{B}_N$  denote the set of multi-sets of intervals (in  $\mathcal{I}$ )  $A$  with  $|A| \leq N$ .

The main result on  $\mathcal{B}_N$  is the following theorem, proved below. (Similar results can also be found in [18].)

**Theorem 3.2.** *For each  $N \geq 0$ ,  $\mathcal{B}_N$  is complete and separable under the bottleneck metric.*

Since the homology  $H_k$  (with any coefficients) of any complex with  $n$  vertices can have rank at most  $\binom{n}{k+1}$ , our persistent homology barcodes will always land in one of the  $\mathcal{B}_N$ , with  $N$  depending just on the size of the samples. As we let the size of the samples increase,  $N$  may increase, and so it is convenient to have a target independent of the number of samples. The space  $\mathcal{B} = \bigcup \mathcal{B}_N$  is clearly not complete under the bottleneck metric, so we introduce the following space of barcodes  $\bar{\mathcal{B}}$ .

**Definition 3.3.** Let  $\bar{\mathcal{B}}$  be the space of multi-sets  $A$  of intervals (in  $\mathcal{I}$ ) with the property that for every  $\epsilon > 0$  the cardinality of the multi-subset of  $A$  of those intervals of length more than  $\epsilon$  has finite cardinality.

Clearly barcodes in  $\overline{\mathcal{B}}$  have at most countable cardinality, and the bottleneck metric extends to a function  $d_{\mathcal{B}}: \overline{\mathcal{B}} \times \overline{\mathcal{B}} \rightarrow \mathbb{R}$ . A straightforward greedy argument shows that for  $X, Y \in \overline{\mathcal{B}}$ ,  $d_{\mathcal{B}}(X, Y) = 0$  only if  $X = Y$ , and so  $d_{\mathcal{B}}$  extends to a metric on  $\overline{\mathcal{B}}$ . We prove the following theorem.

**Theorem 3.4.**  *$\overline{\mathcal{B}}$  is the completion of  $\mathcal{B} = \bigcup \mathcal{B}_N$  in the bottleneck metric. In particular  $\overline{\mathcal{B}}$  is complete and separable in the bottleneck metric.*

*Proof of Theorems 3.2 and 3.4.* The multi-sets of intervals with rational endpoints provides a countable dense subset for  $\mathcal{B}_N$ . To see that  $\mathcal{B}$  is dense in  $\overline{\mathcal{B}}$ , given  $A$  in  $\overline{\mathcal{B}}$  and  $\epsilon > 0$ , let  $A_\epsilon$  be the multi-subset of  $A$  of those intervals of length  $> \epsilon$ . Then by definition of  $\overline{\mathcal{B}}$ ,  $A_\epsilon$  is in  $\mathcal{B}$ , and by definition of the bottleneck metric, using the matching coming from the inclusion of  $A_\epsilon$  in  $A$ , we have that

$$d_{\mathcal{B}}(A, A_\epsilon) \leq \epsilon/2 < \epsilon.$$

It just remains to prove completeness of  $\mathcal{B}_N$  and  $\overline{\mathcal{B}}$ . For this, given a Cauchy sequence  $\langle X_n \rangle$  in  $\mathcal{B}$  it suffices to show that  $X_n$  converges to an element  $X$  in  $\overline{\mathcal{B}}$  and that  $X$  is in  $\mathcal{B}_N$  if all the  $X_n$  are in  $\mathcal{B}_N$ .

Let  $\langle X_n \rangle$  be a Cauchy sequence in  $\mathcal{B}$ . By passing to a subsequence if necessary, we can assume without loss of generality that for  $n, m > k$ ,  $d_{\mathcal{B}}(X_m, X_n) < 2^{-(k+2)}$ . For each  $n$ , we have  $d_{\mathcal{B}}(X_n, X_{n+1}) < 2^{-(n+1)}$ ; choose a matching  $C_n$  such that  $d_\infty(I, J) < 2^{-(n+1)}$  for all  $(I, J) \in C$ . For each  $n$ , define a finite sequence of intervals  $I_1^n, \dots, I_{k_n}^n$  inductively as follows. Let  $k_0 = 0$ . Let  $k_1$  be the cardinality of the multi-subset of  $X_1$  consisting of those intervals of length  $> 1$ , and let  $I_1^1, \dots, I_{k_1}^1$  be an enumeration of those intervals. By induction,  $I_1^n, \dots, I_{k_n}^n$  is an enumeration of the intervals in  $X_n$  of length  $> 2^{-n+1}$  such that for  $i \leq k_{n-1}$ , the intervals  $I_i^{n-1}$  and  $I_i^n$  correspond under the matching  $C_{n-1}$ . For the inductive step, we note that if  $I_i^n$  corresponds to  $J$  under  $C_n$ , then  $d_\infty(I_i^n, J) < 2^{-(n+1)}$ , so the length  $\|J\|$  of  $J$  is bigger than  $\|I_i^n\| - 2^{-n}$ , and

$$\|J\| > 2^{-n+1} - 2^{-n} = 2^{-n} = 2^{-(n+1)+1}.$$

Thus, we can choose  $I_i^{n+1}$  to be the corresponding interval  $J$  for  $i \leq k_n$ , and we can choose the remaining intervals of length  $> 2^{-(n+1)+1}$  in an arbitrary order. Write  $I_i^n = [a_i^n, b_i^n]$  and let

$$a_i = \lim_{n \rightarrow \infty} a_i^n, \quad b_i = \lim_{n \rightarrow \infty} b_i^n.$$

Since  $|a_i^n - a_i^{n+1}| < 2^{-(n+1)}$  and  $|b_i^n - b_i^{n+1}| < 2^{-(n+1)}$ , we have

$$|a_i^n - a_i| \leq 2^{-n}, \quad |b_i^n - b_i| \leq 2^{-n}.$$

Let  $X$  be the multi-subset of  $\mathcal{I}$  consisting of the intervals  $I_i = [a_i, b_i]$  for all  $i$  (or for all  $i \leq \max k_n$  if  $\{k_n\}$  is bounded).

First, we claim that  $X$  is in  $\overline{\mathcal{B}}$ . Given  $\epsilon > 0$ , choose  $N$  large enough that  $2^{-N+2} < \epsilon$ . Then for  $i > k_N$ , the interval  $I_i$  first appears in  $X_{n_i}$  for some  $n_i > N$ . Looking at the matchings  $C_N, \dots, C_{n_i-1}$ , we get a composite matching  $C_{N, n_i}$  between  $X_N$  and  $X_{n_i}$ . Since each  $C_n$  satisfied the bound  $2^{-(n+1)}$ , the matching  $C_{N, n_i}$  must satisfy the bound

$$\sum_{n=N}^{n_i-1} 2^{-(n+1)} = 2^{-N} - 2^{-n_i}.$$

Since all intervals of length  $> 2^{-N+1}$  in  $X_N$  appear as an  $I_j^N$ , we must have that the length of  $I_i^{n_i}$  in  $X_{n_i}$  must be less than

$$2^{-N+1} + 2(2^{-N} - 2^{-n_i}) = 2^{-N+2} - 2^{-n_i+1}.$$

Since each endpoint in  $I_i$  differs from the endpoint of  $I_i^{n_i}$  by at most  $2^{-n_i}$ , the length of  $I_i$  can be at most

$$2^{-N+2} - 2^{-n_i+1} + 2 \cdot 2^{-n_i} = 2^{-N+2} < \epsilon.$$

Thus, the cardinality of the multi-subset of  $X$  of those intervals of length  $> \epsilon$  is at most  $k_N$ .

Next we claim that  $\langle X_n \rangle$  converges to  $X$ . We have a matching of  $X_n$  with  $X$  given by matching the intervals  $I_1^n, \dots, I_{k_n}^n$  in  $X_n$  with the corresponding intervals  $I_1, \dots, I_{k_n}$  in  $X$ . Our estimates above for  $|a_i^n - a_i|$  and  $|b_i^n - b_i|$  show that  $d_\infty(I_i^n, I_i) \leq 2^{-n}$ . By construction, each leftover interval in  $X_n$  has length  $\leq 2^{-n+1}$  and the previous paragraph shows that each leftover interval in  $X$  has length  $< 2^{-n+2}$ . Thus,  $d_B(X_n, X) < 2^{-n+1}$ .

Finally we note that if each  $X_n$  is in  $\mathcal{B}_N$  for fixed  $N$ , then each  $k_n \leq N$  and so  $X$  is in  $\mathcal{B}_N$ .  $\square$

#### 4. FAILURE OF ROBUSTNESS

Inevitably physical measurements will result in bad samples. As a consequence, we are interested in invariants which have limited sensitivity to a small proportion of arbitrarily bad samples. Many standard invariants not only have high sensitivity to a small proportion of bad samples, but in fact have high sensitivity to a small number of bad samples. We use the following terminology.

**Definition 4.1.** A function  $f$  from the set of finite metric spaces to  $\mathbb{R}$  is *fragile* if there exists a constant  $k$  such that for every non-empty finite metric space  $X$  and constant  $N$  there exists a metric space  $X'$  and an isometry  $X \rightarrow X'$  such that  $|X'| \leq |X| + k$  and  $|f(X') - f(X)| > N$ .

Informally, fragile in Definition 4.1 means that adding a small constant number of points to any metric space can arbitrarily change the value of the invariant. In particular, a fragile function is not robust in the sense of Definition 1.4 for any robustness coefficient  $r > 0$ , but fragile is much more unstable than just failing to be robust (note the quantifier on the space  $X$ ). As we indicated in the introduction, Gromov-Hausdorff distance is not robust; here we show it is fragile.

**Proposition 4.2.** Let  $(Z, d_Z)$  be a non-empty finite metric space. The function  $d_{GH}(Z, -)$  is fragile.

*Proof.* Given  $N > 0$ , consider the space  $X'$  which is defined as a set to be the disjoint union of  $X$  with a new point  $w$ , and made a metric space by setting

$$\begin{aligned} d(w, x) &= \alpha, & x \in X, \\ d(x_1, x_2) &= d_X(x_1, x_2), & x_1, x_2 \in X, \end{aligned}$$

where  $\alpha > \text{diam}(Z) + 2d_{GH}(Z, X) + 2N$ . We claim

$$|d_{GH}(Z, X) - d_{GH}(Z, X')| > N.$$

Given any metric space  $(Y, d_Y)$  and isometries  $f: X' \rightarrow Y$ ,  $g: Z \rightarrow Y$ , we need to show that  $d_Y(f(X'), g(Z)) > N + d_{GH}(Z, X)$ . We have two cases. First, if no

point  $z$  of  $Z$  has  $d_Y(g(z), f(w)) \leq N + d_{GH}(Z, X)$ , then we have  $d_Y(f(X'), g(Z)) > N + d_{GH}(Z, X)$ . On the other hand, if some point  $z$  of  $Z$  has  $d_Y(g(z), f(w)) < N + d_{GH}(Z, X)$ , then every point  $z$  in  $Z$  satisfies  $d_Y(g(z), f(w)) \leq N + \text{diam}(Z) + d_{GH}(Z, X)$ . Choosing some  $x$  in  $X$ , we see that for every  $z$  in  $Z$ ,  $d_Y(f(x), g(z)) \geq \alpha - (N + \text{diam}(Z) + d_{GH}(Z, X))$ . It follows that

$$d_Y(f(X'), g(Z)) \geq \alpha - (N + \text{diam}(Z) + d_{GH}(Z, X)) > N + d_{GH}(Z, X). \quad \square$$

The homology and persistent homology of a point cloud turns out to be a somewhat less sensitive invariant. Nonetheless, a similar kind of problem can occur. It is instructive to consider the case of  $H_0$  or  $\text{PH}_0$ . By adding  $\ell$  points far from the original metric space  $X$ , one can change either  $H_0$  or  $\text{PH}_0$  by rank  $\ell$ . The further the distance of the points, the longer the additional bars in the bar code and we see for example that the distance  $d_{\mathcal{B}}(B, -)$  in the bottleneck metric from any fixed bar code  $B$  is a fragile function. (If we are truncating the bar codes,  $d_{\mathcal{B}}$  is bounded by the length of the interval we are considering, so technically is robust, but not in a meaningful way.) We can also consider the rank of  $H_0$  or of  $\text{PH}_0$  in a range; here the distortion of the function depends on the number of points, but we see that the function is not robust.

For  $H_k$  and  $\text{PH}_k$ ,  $k \geq 0$ , the same basic idea obtains: we add small spheres sufficiently far from the core of the points in order to adjust the required homology. We work this out explicitly for  $\text{PH}_1$ .

**Definition 4.3.** For each integer  $k > 0$  and real  $\ell > 0$ , let the metric circle  $S_{k,\ell}^1$  denote the metric space with  $k$  points  $\{x_i\}$  such that

$$d(x_i, x_j) = \ell (\min(|i - j|, |k - i - j|)).$$

For  $\epsilon < \ell$ , the Rips complex associated to  $S_{k,\ell}^1$  is just a collection of disconnected points. It is clear that as long as  $k \geq 4$ , when  $\ell \leq \epsilon < 2\ell$ ,  $|R_\epsilon(S_{k,\ell}^1)|$  has the homotopy type of a circle. In fact, we can say something more precise:

**Lemma 4.4.** *For*

$$\ell \leq \epsilon < \left\lceil \frac{k}{3} \right\rceil \ell,$$

*the rank of  $H_1(R_\epsilon(S_{k,\ell}^1))$  is at least 1.*

*Proof.* Consider the map  $f$  from  $R_\epsilon(S_{k,\ell}^1)$  to the unit disk  $D^2$  in  $\mathbb{R}^2$  that sends  $x_i$  to  $(\cos(2\pi \frac{i}{n}), \sin(2\pi \frac{i}{n}))$  and is linear on each simplex. The condition  $\epsilon < \lceil \frac{k}{3} \rceil \ell$  precisely ensures that whenever  $\{x_{i_1}, \dots, x_{i_n}\}$  forms a simplex  $\sigma$  in the Rips complex, the image vertices  $f(x_{i_1}), \dots, f(x_{i_n})$  lie on an arc of angle  $< \frac{2}{3}\pi$  on the unit circle, and so  $f(\sigma)$  in particular lies in an open half plane through the origin. It follows that the origin  $(0,0)$  is not in the image of any simplex, and  $f$  defines a map from  $R_\epsilon(S_{k,\ell}^1)$  to the punctured disk  $D^2 - \{(0,0)\}$ . Since  $\ell \leq \epsilon$ , we have the 1-cycle

$$[x_1, x_2] + \dots + [x_{k-1}, x_k] + [x_k, x_1]$$

of  $R_\epsilon(S_{k,\ell}^1)$  which maps to a 1-cycle in  $D^2 - \{(0,0)\}$  representing the generator of  $H_1(D^2 - \{(0,0)\})$ .  $\square$

The length  $\ell$  and number  $k \geq 4$  is arbitrary, so again, we conclude that functions like  $d_{\mathcal{B}}(B, \text{PH}_1(-))$  are fragile. Results for higher dimensions (using similar standard equidistributed models of  $n$ -spheres) are completely analogous.

**Proposition 4.5.** *Let  $B$  be a barcode. The functions  $d_{\mathcal{B}}(B, \text{PH}_k(-))$  from finite metric spaces to  $\mathbb{R}$  are fragile.*

In terms of rank, the lemma shows that we can increase the rank of first persistent homology group of a metric space  $X$  on an interval  $[a, b]$  by  $m$  simply by adding extra points. One can also typically reduce persistent homology intervals by adding points “in the center” of the representing cycle. It is somewhat more complicated to precisely analyze the situation, so we give a representative example: Suppose the cycle is represented by a collection of points  $\{x_i\}$  such that the maximum distance  $d(x_i, x_j) \leq \delta$ . Then adding a point which is a distance  $\delta$  from each of the other points reduces the lifetime of that cycle to  $\delta$ . In any case, the results of the lemma are sufficient to prove the following proposition.

**Proposition 4.6.** *The function that takes a finite metric space to the rank of  $\text{PH}_k$  on a fixed interval  $[a, b]$  is not robust (in the sense of Definition 1.4) for any robustness coefficient  $r$ .*

These computations suggest a problem with the stability of the usual invariants of computational topology. A small number of bad samples can lead to arbitrary changes in these invariants.

## 5. THE MAIN DEFINITION AND THEOREM

Fix a good functorial assignment of a simplicial complex to a finite metric space and a scale parameter  $\epsilon$ . Recall that we write  $\text{PH}_k$  of a finite metric space to denote the persistent homology of the associated direct system of complexes. Motivated by the concerns of the preceding section, we make the following definition.

**Definition 5.1.** For a metric measure space  $(X, \partial_X, \mu_X)$  and fixed  $n, k \in \mathbb{N}$ , define the  $k$ th  $n$ -sample persistent homology as

$$\Phi_k^n(X, \partial_X, \mu_X) = (\text{PH}_k)_*(\mu_X^{\otimes n}),$$

the probability distribution on  $\bar{\mathcal{B}}$  induced by pushforward along  $\text{PH}_k$  from the product measure  $\mu_X^n$  on  $X^n$ .

The goal of this section is to prove the main theorem. For this (and in the remainder of the section), we assume that we are computing PH using the Rips complex.

**Theorem 5.2.** *Let  $(X, \partial_X, \mu_X)$  and  $(Y, \partial_Y, \mu_Y)$  be compact metric measure spaces. Then we have the following inequality:*

$$d_{Pr}(\Phi_k^n(X, \partial_X, \mu_X), \Phi_k^n(Y, \partial_Y, \mu_Y)) \leq n d_{GPr}((X, \partial_X, \mu_X), (Y, \partial_Y, \mu_Y)).$$

*Proof.* Assume that  $d_{GPr}((X, \partial_X, \mu_X), (Y, \partial_Y, \mu_Y)) < \epsilon$ . Then we know that there exist embeddings  $\iota_1: X \rightarrow Z$  and  $\iota_2: Y \rightarrow Z$  into a metric space  $Z$  and a coupling  $\hat{\mu}$  between  $(\iota_1)_*\mu_X$  and  $(\iota_2)_*\mu_Y$  such that the probability mass of the set of pairs  $(z, z')$  under  $\hat{\mu}$  such that  $\partial_Z(z, z') \geq \epsilon$  is less than  $\epsilon$ .

We can regard the restriction of  $\hat{\mu}^{\otimes n}$  to the full measure subspace  $(X \times Y)^n$  of  $(Z \times Z)^n$  as a probability measure on  $X^n \times Y^n$ . This then induces a coupling between  $\text{PH}_k(\mu_X^{\otimes n})$  and  $\text{PH}_k(\mu_Y^{\otimes n})$  on  $\bar{\mathcal{B}}$ , which we now study. Consider  $n$  samples  $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$  from  $Z \times Z$  drawn according to the product distribution  $\hat{\mu}^{\otimes n}$ . Now consider the probability that

$$\alpha = \sup_{1 \leq i, j \leq n} |\partial_X(x_i, x_j) - \partial_Y(y_i, y_j)| \geq 2\epsilon.$$

The triangle inequality implies that

$$|\partial_X(x_i, x_j) - \partial_Y(y_i, y_j)| \leq \sup(\partial_Z(x_i, x_j), \partial_Z(y_i, y_j)) \leq \partial_Z(x_i, y_i) + \partial_Z(x_j, y_j).$$

Therefore, the union bound implies that the probability that  $\alpha \geq 2\epsilon$  is bounded by

$$\sum_{i=1}^n \Pr(\partial_Z(x_i, y_i) \geq \epsilon) \leq n\epsilon.$$

Next, define a relation  $R$  that matches  $x_i$  and  $y_i$ . By definition, the distortion of this relation is  $\text{dis } R = \alpha$ , and so

$$d_{GH}(\{x_i\}, \{y_i\}) \leq \frac{1}{2}\alpha.$$

By the stability theorem of Chazal, et. al. [6, 3.1] (Theorem 2.9 above), this implies that the probability that

$$d_{\mathcal{B}}(\text{PH}_k(\{x_i\}), \text{PH}_k(\{y_i\})) \geq \epsilon$$

is bounded by  $n\epsilon$ . This further implies that the probability that

$$d_{\mathcal{B}}(\text{PH}_k(\{x_i\}), \text{PH}_k(\{y_i\})) \geq n\epsilon$$

is also bounded by  $n\epsilon$ . Therefore, we can conclude that

$$d_{Pr}(\Phi_k^n(X, \partial_X, \mu_X), \Phi_k^n(Y, \partial_Y, \mu_Y)) \leq n\epsilon. \quad \square$$

One of the consequences of Theorem 5.2 is that  $\Phi_k^n$  provide robust descriptors for metric measure spaces  $(X, \partial_X, \mu_X)$ . Specifically, observe that if we have  $(X, \partial_X) \subset (X', \partial_{X'})$  and a probability measure  $\mu_{X'}$  that restricts to  $\mu_X$  on  $X$ , then

$$d_{Pr}(i_*\mu_X, \mu_{X'}) \leq 1 - \mu_{X'}(X).$$

Thus, when  $X' \setminus X$  has probability  $< \epsilon$ ,

$$d_{Pr}(\Phi_k^n(X, \partial_X, \mu_X), \Phi_k^n(X', \partial_{X'}, \mu_{X'})) \leq n\epsilon.$$

In particular, when  $X$  and  $X'$  are finite metric spaces with the uniform measure, we get

$$d_{Pr}(\Phi_k^n(X, \partial_X, \mu_X), \Phi_k^n(X', \partial_{X'}, \mu_{X'})) \leq n(1 - |X|/|X'|).$$

As an immediate consequence we obtain the following result.

**Theorem 5.3.** *For fixed  $n, k$ ,  $\Phi_k^n$  is uniformly robust with robustness coefficient  $r$  and estimate bound  $nr/(1+r)$  for any  $r$ .*

To study more general metric measure spaces, the Glivenko-Cantelli theorem implies that consideration of large finite samples will suffice.

**Lemma 5.4.** *Let  $S_1 \subset S_2 \subset \dots \subset S_i \subset \dots$  be a sequence of randomly drawn samples from  $(X, \partial_X, \mu_X)$ . We regard  $S_i$  as a metric measure space using the subspace metric and the empirical measure. Then  $\Phi_k^n(S_i)$  converges almost surely to  $\Phi_k^n((X, \partial_X, \mu_X))$ .*

*Proof.* This is a consequence of the fact that  $\{S_i\}$  converges almost surely to  $(X, \partial_X, \mu_X)$  in the Gromov-Prohorov metric (which can be checked directly or for instance follows from the analogous result for the Gromov-Wasserstein distance [22, 3.5.(iii)] and the comparison between the Gromov-Prohorov distance and the Gromov-Wasserstein distance [11, 10.5].  $\square$

*Remark 5.5.* It would be useful to prove analogues of the main theorem for other methods of assigning complexes; e.g., the witness complex (see Remark 2.4 and Section 8).

## 6. HYPOTHESIS TESTING, CONFIDENCE INTERVALS, AND NUMERICAL INVARIANTS

In this section, we discuss hypothesis testing and the closely related issue of confidence intervals for our barcode distribution invariants. The basic goal is to provide quantitative ways of saying what an observed empirical barcode distribution means. The model for hypothesis testing is that we have two empirical distributions (obtained from some kind of sampling) and we want to estimate the probability that they were drawn from the same distribution. We demonstrate how to do this for  $\Phi_k^n$  in some synthetic examples in the next section. In Section 8, we apply this to validate part of the analysis of the natural images dataset in [3].

In our setting we cannot assume anything about the class of possible hypotheses and so we are forced to rely on non-parametric methods. Most results on non-parametric tests for distribution comparison work only for distributions on  $\mathbb{R}$  and the first step is to project the data into this framework. In our examples, we use two kinds of projections.

**Definition 6.1.** Let  $(X, \partial_X, \mu_X)$  be a compact metric measure space. Fix  $k, n \in \mathbb{N}$ . Define the distance distribution  $\mathcal{D}^2$  on  $\mathbb{R}$  to be the distribution on  $\mathbb{R}$  induced by applying  $d_B(-, -)$  to pairs  $(b_1, b_2)$  drawn from  $\Phi_k^n(X, \partial_X, \mu_X)^{\otimes 2}$ . Let  $B$  be a fixed barcode in  $\bar{\mathcal{B}}$ , and define  $\mathcal{D}_B$  to be the distribution induced by applying  $d_B(B, -)$ .

We will demonstrate hypothesis testing for comparison of distributions using these projections and the two-sample Kolmogorov-Smirnov statistic [8, §6]. This test gives a way to determine whether two observed empirical distributions were obtained from the same underlying distribution; moreover, for distributions on  $\mathbb{R}$  the  $p$ -values of the test statistic are independent of the underlying distribution as long as the samples are identically independently drawn.

To compute the Kolmogorov-Smirnov test statistic for two sets of samples  $\mathcal{S}_1$  and  $\mathcal{S}_2$ , we first compute the empirical approximations  $\mathcal{E}_1$  and  $\mathcal{E}_2$  to the cumulative density functions,

$$E_i(t) = |\{x \in S_i \mid x \leq t\}|/|S_i|,$$

and use the test statistic  $\sup_t |\mathcal{E}_1(t) - \mathcal{E}_2(t)|$ . (In practice,  $|S_i|$  is large and we approximate  $\mathcal{E}_i$  using Monte Carlo methods.) Standard tables (e.g., in the appendix to [8]) or the built-in Matlab functions can then be used to compute  $p$ -values for deciding if the statistic allows us to reject the hypothesis that the distributions are the same.

*Remark 6.2.* Our choice of the Kolmogorov-Smirnov test was arbitrary and for purposes of illustration; one might also consider the Mann-Whitney test or various other nonparametric techniques for testing whether samples were drawn from the same underlying distribution.

As a second demonstration of hypothesis testing, we can apply a  $\chi^2$  test to discrete distributions constructed by the following procedure. Fix a finite set  $\{B_j\} \subset \bar{\mathcal{B}}$  of reference barcodes, where  $1 \leq j \leq m$ . For each barcode with nonzero probability measure in (the given empirical approximation to)  $\Phi_k^n$ , assign the count to the nearest reference barcode. (In general, we compute the empirical approximation to



$\Phi_k^n$  using Monte Carlo methods.) Let  $\mathcal{A}_i(j)$  denote the count for reference barcode  $B_j$  in sample  $i$  (for  $i = 1, 2$ ). The test statistic in the test is defined to be

$$\chi^2 = \sum_{j=1}^m \frac{(\mathcal{A}_1(j) - \mathcal{A}_2(j))^2}{\mathcal{A}_1(j) + \mathcal{A}_2(j)}.$$

As the notation suggest, asymptotically this has a  $\chi^2$  distribution with  $m' - 1$  degrees of freedom (where  $m'$  is the number of reference barcodes with nonzero counts). As such, we can again look up the  $p$ -values for this distribution in standard tables.

*Remark 6.3.* Independence of distribution for  $p$ -values above hold asymptotically as a consequence of the Glivenko-Cantelli theorem; however, in general we may worry if our sample sizes are large enough for the  $p$ -value to be correct.

Another approach it to use summary test statistics. A very natural test statistic associated to  $\Phi_k^n$  measures the distance to a fixed hypothesis distribution.

**Definition 6.4.** Let  $(X, \partial_X, \mu_X)$  be a compact metric measure space and let  $\mathcal{P}$  be a fixed reference distribution on  $\bar{\mathcal{B}}$ . Fix  $k, n \in \mathbb{N}$ . Define the homological distance on  $X$  relative to  $\mathcal{P}$  to be

$$\text{HD}_k^n((X, \partial_X, \mu_X), \mathcal{P}) = d_{Pr}(\Phi_k^n(X, \partial_X, \mu_X), \mathcal{P}).$$

The proof of Lemma 5.4 applies to show that large finite samples suffice to approximate  $\text{HD}_k^n$ .

**Lemma 6.5.** *Let  $S_1 \subset S_2 \subset \dots \subset S_i \subset \dots$  be a sequence of randomly drawn samples from  $(X, \partial_X, \mu_X)$ . We regard  $S_i$  as a metric measure space using the subspace metric and the empirical measure. Then for  $\mathcal{P}$  a fixed reference distribution on  $\bar{\mathcal{B}}$ ,  $\text{HD}_k^n(S_i, \mathcal{P})$  converges almost surely to  $\text{HD}_k^n((X, \partial_X, \mu_X), \mathcal{P})$ .*

An immediate consequence of Theorem 5.2 is the following robustness result (paralleling Theorem 5.3).

**Theorem 6.6.** *For fixed  $n, k, \mathcal{P}$ ,  $\text{HD}_k^n(-, \mathcal{P})$  is uniformly robust with robustness coefficient  $r$  and estimate bound  $nr/(1+r)$  for any  $r$ .*

Alternatively, a virtue of distributions on  $\mathbb{R}$  is that they can be naturally summarized by moments; in contrast, moments for distributions on barcode space are very hard to understand (for example, geodesics between close points are not unique). The first moment is the mean, which could be used as a test statistic. Because we have emphasized robust statistics, we work instead with the median (or a trimmed mean, see Remark 6.9) and introduce the following test statistic:

**Definition 6.7.** Let  $(X, \partial_X, \mu_X)$  be a compact metric measure space and let  $\mathcal{P}$  be a fixed reference barcode  $B \in \mathcal{B}$ . Fix  $k, n \in \mathbb{N}$ . Let  $\mathcal{D}$  denote the distribution on  $\mathbb{R}$  induced by applying  $d_B(B, -)$  to the barcode distribution  $\Phi_k^n(X, \partial_X, \mu_X)$ . Define the median homological distance relative to  $\mathcal{P}$  to be

$$\text{MHD}_k^n((X, \partial_X, \mu_X), \mathcal{P}) = \text{median}(\mathcal{D}).$$

Again, the Glivenko-Cantelli theorem implies that consideration of large finite samples will suffice.

**Lemma 6.8.** *Let  $S_1 \subset S_2 \subset \dots \subset S_i \subset \dots$  be a sequence of randomly drawn samples from  $(X, \partial_X, \mu_X)$ . We regard  $S_i$  as a metric measure space using the subspace metric and the empirical measure. Let  $\mathcal{P}$  be a fixed reference barcode, and assume that  $\mathcal{D}((X, \partial_X, \mu_X), \mathcal{P})$  has a distribution function with a positive derivative at the median. Then  $\text{MHD}_k^n(S_i, \mathcal{P})$  almost surely converges to  $\text{MHD}_k^n((X, \partial_X, \mu_X), \mathcal{P})$ .*

*Proof.* As in the proof of Lemma 5.4, the fact that  $\{S_i\}$  converges almost surely to  $(X, \partial_X, \mu_X)$  in the Gromov-Prohorov metric implies that  $\mathcal{D}(S_i)$  weakly converges to  $\mathcal{D}$ . Now the central limit theorem for the sample median (see for instance [20, III.4.24]) implies the convergence of medians.  $\square$

**Remark 6.9.** To remove the (possibly hard to verify) hypothesis in Lemma 6.8, one can replace the median with a trimmed mean (i.e., the mean of the distribution obtained by throwing away the top and bottom  $k\%$ , for some constant  $k$ ).

As discussed in the introduction, a counting argument yields the following robustness result.

**Theorem 6.10.** *For any  $n, k, \mathcal{P}$ , the function  $\text{MHD}_k^n(-, \mathcal{P})$  from finite metric spaces (given the uniform probability measure) to  $\mathbb{R}$  is robust with robustness coefficient  $> (\ln 2)/n$ .*

A particular advantage of  $\text{HD}_k^n$  and  $\text{MHD}_k^n$  is that we can define confidence intervals using the standard non-parametric techniques for determining confidence intervals for the median [9, §7.1]. Specifically, we use appropriate sample quantiles (order statistics) to determine the bounds for an interval which contains the actual median with confidence  $1 - \alpha$ . For example, a simple approximation can be obtained from the fact that order statistics asymptotically obey binomial distributions, which lead to the following definition using the normal approximation to the binomial distribution.

**Definition 6.11.** Let  $(X, \partial_X, \mu_X)$  be a metric measure space,  $\mathcal{P}$  a fixed barcode, and  $\tilde{\mathcal{P}}$  a fixed distribution on barcodes. Fix  $0 \leq \alpha \leq 1$  and  $n, k$ . Given  $m$  empirical approximations to  $\text{HD}_k^n(-, \tilde{\mathcal{P}})$  or  $\text{MHD}_k^n(-, \mathcal{P})$ , let  $\{s_m\}$  denote the samples sorted from smallest to largest. Let  $u_\alpha$  denote the  $\frac{\alpha}{2}$  significance threshold for a standard normal distribution. The  $1 - \alpha$  confidence interval for the sample median is given by the interval

$$\left[ s_{\lfloor \frac{m+1}{2} - \frac{1}{2} \sqrt{m} u_\alpha \rfloor}, s_{\lceil \frac{m+1}{2} + \frac{1}{2} \sqrt{m} u_\alpha \rceil} \right].$$

Note that a further advantage of  $\text{HD}_k^n$  and  $\text{MHD}_k^n$  is that the asymptotics we rely on for the confidence intervals in the preceding definition do not depend on the convergence result in the main theorem, and as a consequence we can apply these confidence intervals when studying filtered complexes generated by a procedure other than the Rips complex, e.g., the witness complex. We give examples in Section 8 of both sorts of confidence interval.

More sophisticated estimates involving better techniques for non-parametric confidence intervals for the population median can also be adopted. Of course, such non-parametric estimates depend on the convergence to the binomial approximation, which we cannot know a priori (cf. Remark 6.3 above). Furthermore, this requires empirical estimates of the distribution  $\Phi_k^n(X, \partial_X, \mu_X)$ . But provided this is available, we can perform confidence-interval based hypothesis testing.

*Remark 6.12.* All of the inference procedures described in this section require repeated sampling of samples of size  $k$ . This raises the question that, given the opportunity to obtain  $m$  samples of size  $k$ , is it more sensible to regard this as  $m$  samples of size of size  $k$  or one sample of size  $mk$  (or something in between)? Considerations of the latter sort quickly lead to consideration of bootstrap and re-sampling methods; we believe that bootstrap confidence intervals are most likely to be useful in this setting. We study resampling, proving asymptotic consistency and studying the convergence behaviors in the companion paper [1].

Finally, in principle we would like to be able to use the distribution invariant  $\Phi_k^n(X, \partial_X, \mu_X)$  (or perhaps other distributions of barcodes) in likelihood statistics. For example, given a hypothesis barcode  $B$ , we can empirically estimate the probability that one would see a barcode within  $\epsilon$  of  $B$  when sampling  $n$  points. This then in principle allows us to test whether a particular sample (or collection of samples) is consistent with the hypothesis  $\text{Hyp}_k^n(X; B, \epsilon)$  that  $B$  is within  $\epsilon$  of the barcode of a sample of size  $n$  drawn from  $(X, \partial_X, \mu_X)$ . More generally, we can distinguish between  $(X, \partial_X, \mu_X)$  and  $(X', \partial_{X'}, \mu_{X'})$  by calculating likelihoods of the hypotheses  $\text{Hyp}_k^n(X; B, \epsilon)$  and  $\text{Hyp}_k^n(X'; B, \epsilon)$ . Specifically, given an observed barcode  $B$  obtained by sampling  $n$  points from an unknown metric measure space  $(Z, \partial_Z, \mu_Z)$ , we can compute the likelihood

$$L_X = L(X, \partial_X, \mu_X) = \Pr(d_B(B, \tilde{B}) < \epsilon \mid \tilde{B} \text{ drawn from } \Phi_k^n(X, \partial_X, \mu_X)).$$

The ratio  $L_X/L_{X'}$  provides a test statistic for comparing the two hypotheses. A significant difficulty with this approach, however, is that in order to compute the thresholds on  $L_X/L_{X'}$  for deciding which hypothesis to accept at a given  $p$ -value, we require knowledge of the distribution of the likelihood scores induced by  $\Phi_k^n(X, \partial_X, \mu_X)$  and  $\Phi_k^n(X', \partial_{X'}, \mu_{X'})$ , which in general must be obtained by Monte Carlo simulation (i.e., repeated sampling).

We can overcome the difficulty above by instead testing the likelihood of a more distributional statement. For a metric measure space  $(X, \partial_X, \mu_X)$  and a subset  $S$  of  $\overline{B}$ , we can estimate the likelihood that the distribution  $\Phi_k^n(X)$  has mass  $\geq \epsilon$  on  $S$  as follows. For any hypothetical distribution on  $\overline{B}$  with mass  $\geq \epsilon$  on  $S$ , the probability of an empirical sample of size  $N$  having  $q$  or fewer elements in  $S$  is bounded above by the binomial cumulative distribution function

$$\text{BD}(N, q, \epsilon) = \sum_{i=0}^q \binom{N}{i} \epsilon^i (1 - \epsilon)^{N-i}.$$

Then given an empirical approximation  $\mathcal{E}$  to  $\Phi_k^n$  obtained from  $N$  samples, we can test the hypothesis that  $\Phi_k^n$  has mass  $\geq \epsilon$  in  $S$ , by taking  $q$  to be the number of such elements in  $\mathcal{E}$ . When  $\text{BD}(N, q, \epsilon) < \alpha$ , we can reject this hypothesis at the  $1 - \alpha$  level.

*Remark 6.13.* In the test statistics in this section, we have suggested using Monte Carlo methods to estimate distributions  $\Phi_k^n$ , but not to estimate  $p$ -values. The reason is that the space of possible null hypotheses is often so large that it is infeasible to imagine doing the required Monte Carlo simulations.

## 7. DEMONSTRATION OF HYPOTHESES TESTING ON SYNTHETIC EXAMPLES

**Synthetic Example 1: The annulus and the annulus plus diameter lineage.** To demonstrate hypothesis testing methodology for  $\Phi_k^n$ , we first consider a simple

example which illustrates the robustness of the distributional invariants. We generated a set  $\mathcal{S}_1$  of 1000 points by sampling uniformly (via rejection sampling) from an annulus of inner radius 0.8 and outer radius of 1.2 in  $\mathbb{R}^2$  (see Figure 1). The underlying manifold is homotopy equivalent to a circle, and computing the barcode for the first homology group (with cutoff of 0.75) yields a single long interval, displayed in Figure 2.

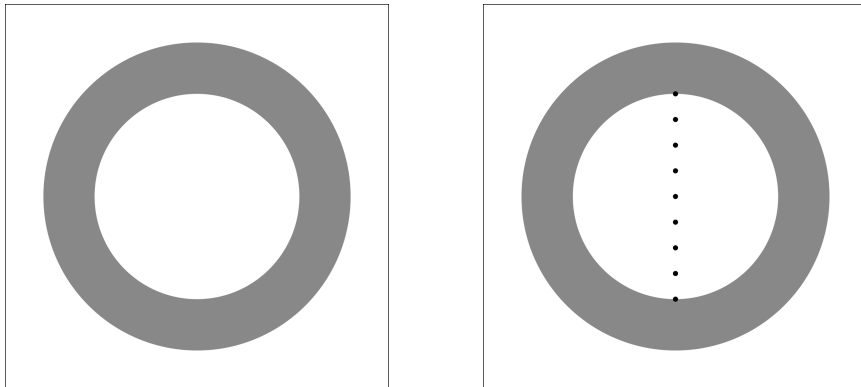


FIGURE 1. Annulus with inner radius 0.8 and outer radius 1.2 (left), and same annulus together with diameter linkage given by the nine points  $(0, 0.8)$ ,  $(0, 0.6)$ ,  $(0, 0.4)$ ,  $(0, 0.2)$ ,  $(0, 0)$ ,  $(0, -0.2)$ ,  $(0, -0.4)$ ,  $(0, -0.6)$ ,  $(0, -0.8)$  (right).

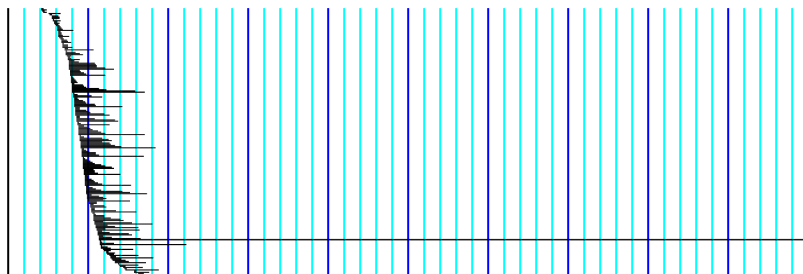


FIGURE 2. Barcode for annulus via the Rips complex with 1000 points. Horizontal scale goes from 0 to 0.75. (Vertical scale is not meaningful.)

We then added the set of points  $X$

$$\{(0, 0.8), (0, 0.6), (0, 0.4), (0, 0.2), (0, 0), (0, -0.2), (0, -0.4), (0, -0.6), (0, -0.8)\}$$

to form the set  $\mathcal{S}_2 = \mathcal{S}_1 \cup X$ . By adding these 9 points, the point cloud now appears to have been sampled from an underlying manifold homotopy equivalent to a figure 8 when the parameter is between 0.2 and our cutoff 0.75. (See Figure 1.)

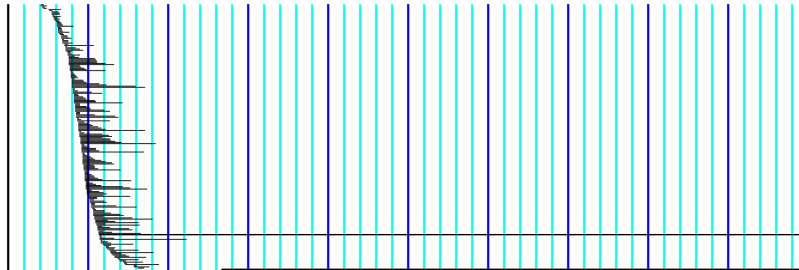


FIGURE 3. Barcode for the annulus plus diameter linkage via the Rips complex with 1000 points. Horizontal scale goes from 0 to 0.75. (Vertical scale is not meaningful.)

Computing the barcode for the first homology group now yields two long intervals, displayed in Figure 3.

We computed empirical approximations to  $\Phi_1^{75}(\mathcal{S}_1)$  and  $\Phi_1^{75}(\mathcal{S}_2)$ , using 1000 samples of size 75 and barcode cutoffs of 0.75. Comparing the empirical distance distributions  $\mathcal{D}^2$  (as in Definition 6.1) using the Kolmogorov-Smirnov statistic suggested they were drawn from the same distribution at the 95% confidence level. Fixing a reference barcode  $B_1$  with a single long bar and using the associated distance distribution produced the same result.

For the  $\chi^2$  test on these samples, we found that the resulting distributions had nontrivial mass clustered in three regions: around a barcode  $B_0$  with no long intervals, a barcode  $B_1$  with one long interval, and a barcode  $B_2$  with two long intervals. Assigning each point in the empirical approximation to the nearest such barcode, the distribution of masses were 0.236, 0.749, and 0.015 for  $\mathcal{S}_1$  and 0.234, 0.741, and 0.025 for  $\mathcal{S}_2$ . The  $\chi^2$  test suggested they were drawn from the same distribution at the 95% confidence level.

This example also begins to illuminate a relationship between the distributional invariants and density filtering. Notice that the second interval at the bottom of Figure 3 starts somewhat later, reflecting a difference in average interpoint distance between the original samples and the additional points added. As a consequence, one might imagine that appropriate density filtering would also detect these points. On the one hand, in many cases density filtering is an excellent technique for concentrating on regions of interest. On the other hand, it is easy to construct examples where density filtering fails (for instance, we can build examples akin to the one studied here where the “connecting strip” has comparable density to the rest of the annulus simply by reducing the number of sampled points or by expanding the outer radius while keeping the number of sampled points fixed). More generally, studying distributional invariants (such as  $\Phi$ ) by definition allows us to integrate information from different density scales. In practice, we expect there to be a synergistic interaction between density filtering and the use of  $\Phi_k^n$ ; see Section 8 for an example.

**Synthetic Example 2: Friendly circles.** Next, we considered a somewhat more complicated example. We sampled 750 points uniformly in the volume measure from

the circle centered at  $(0,0)$  of radius 2, and 750 points uniformly in the volume measure from the circle centered at  $(0.8,0)$  of radius 1. We then added uniform noise sampled from  $[-2,2] \times [-2,2] \subset \mathbb{R}^2$ . (See Figure 4.) Without noise, we saw the expected pair of long bars in the barcode for the first persistent homology group, computed using the Rips complex. However, as noise was added, the results of computing barcodes using the Rips complex degraded very rapidly, as we see in Figure 5.

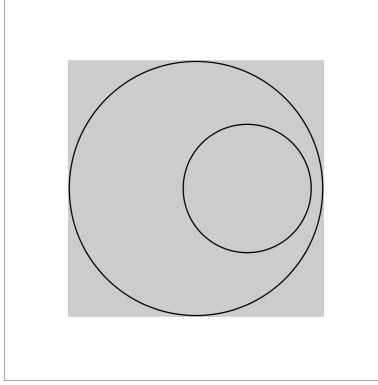


FIGURE 4. Two circles with noise (indicated by gray box).

Even with only 10 noise points, we see 3 bars, and with 90 noise points there are 12. (These results were stable across different samples; we report results for a representative run.)

We computed  $\Phi_1^{300}$  for the same point clouds (i.e., the two circles plus varying numbers of noise points), using 1000 samples of size 300 and a cutoff of 0.75. The resulting empirical distributions had essentially all of their weight concentrated around barcodes with a small number of long intervals. We clustered the points in the empirical estimate of  $\Phi_1^{300}$  around barcodes with fixed numbers of long bars (from 0 to the cutoff), assigning each point to the nearest barcode. The results are summarized in Table 1 below.

TABLE 1. Distribution summaries for  $\Phi_1^{300}$

Number of noise pts	0 bars	1 bars	2 bars	3 bars	4 bars	5 bars
0	0	303	696	1	0	0
10	0	305	589	106	0	0
20	0	278	590	132	0	0
30	0	285	594	119	2	0
40	1	259	584	149	6	1
50	0	289	553	154	4	0
60	0	254	591	146	7	2
70	0	277	564	154	5	0
80	1	229	543	196	29	2
90	0	229	533	207	28	3

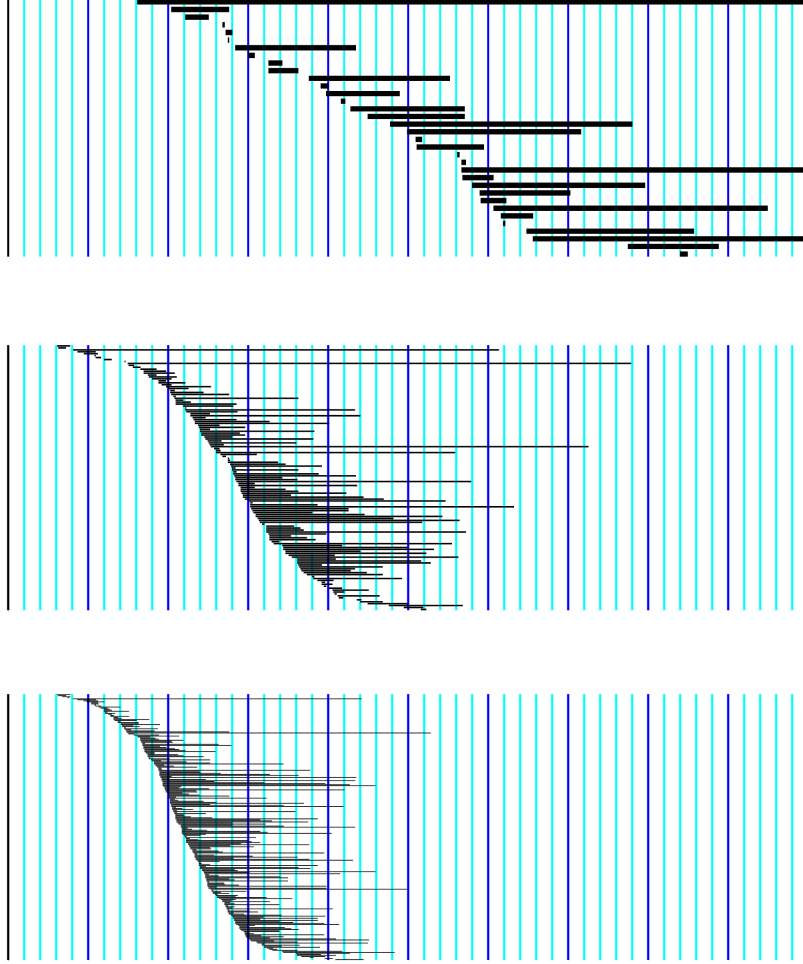


FIGURE 5. Barcode for two circles with 10, 50, and 90 noise points. Horizontal scale goes from 0 to 0.75. (Vertical scale is not meaningful.)

A glance at the table shows that the majority of the weight is clustered around a barcode with 2 long bars and that the data overwhelmingly supports a hypothesis of  $\leq 3$  barcodes under all noise regimes. We can be more precise using the likelihood statistic of the last section to evaluate the hypothesis  $H$  that the observed empirical approximation to  $\Phi_k^n$  was drawn from an underlying barcode distribution with weight  $\geq 5\%$  on barcodes with more than 3 long bars. In the strictest tests with 80 and 90 noise points, 31 out of 1000 samples were near bar codes with more than 3 long bars, and so we estimate that the probability of the distribution having  $\geq 5\%$  of the mass at 4 or more bar codes as  $\leq \text{BD}(1000, 31, .05) < 0.22\%$ . Put another

way, we can reject the hypothesis that the actual distribution has more than 5% mass at 4 or more bar codes at the 99.7% level.

## 8. APPLICATION: CONFIDENCE INTERVALS FOR THE NATURAL IMAGES DATASET

**8.1. Setup.** In this section, we compute the confidence intervals based on  $\text{MHD}_k^n$  for a subset of patches from the natural images data set as described in [3]. We briefly review the setup. The dataset consists of 15000 points in  $\mathbb{R}^8$ , generated as follows. From the natural images,  $3 \times 3$  patches (dimensions given in pixels) were sampled and the top 20% with the highest contrast were retained. These patches were then normalized twice, first by subtracting the mean intensity and then scaling so that the Euclidean norm is 1. The resulting dataset can be regarded as living on the surface of an  $S^7$  embedded in  $\mathbb{R}^8$ . After performing density filtering (with a parameter value of  $k = 15$ ; refer to [3] for details) and randomly selecting 15000 points, we are left with the dataset  $\mathcal{M}(15, 30)$ . At this density, one tends to see a barcode corresponding to 5 cycles in the  $H_1$ . In the Klein bottle model, these cycles are generated by three circles, intersecting pairwise at two points (which can be visualized as unit circles lying on the  $xy$ -plane, the  $yz$ -plane, and the  $xz$ -plane).

**8.2. Results.** We computed an empirical approximation to  $\Phi_1^{500}(\mathcal{M}(15, 30))$  and found that (after clustering, as above) the weight was distributed as 0.1 % with one bar, 1.1 % with two bars, 7.4% with three bars, 34.2% with four bars, and 57.2% with five bars. Analyzing likelihoods, we can see that the underlying distribution has at least 95% of its mass on two, three, or four bars at the the 99.7% confidence level.

We also analyze the results using MHD. We use as the hypothesis barcode the multi-set  $\{(a, c), (a, c), (a, c), (a, c), (a, c)\}$ , where  $a$  is chosen to be smaller than the minimum value at which any signal appears in the first Betti number and  $c$  is the maximum bar length found in the dataset (and is typically an arbitrary value obtained as the cutoff for the maximal filtration value; we used the value reported in [3], which is 2). We approximate  $\text{MHD}_1^{500}(\mathcal{M}(15, 30))$  by the empirical distribution based on 1000 samples. We find using the non-parametric estimate based on the interval statistics that the 95% confidence interval for  $\text{MHD}_1^{500}(\mathcal{M}(15, 30))$  is  $[0.442, 0.476]$ . The 99% confidence interval for  $\text{MHD}_1^{500}(\mathcal{M}(15, 30))$  is  $[0.436, 0.481]$ . These results represents high confidence for the data to be further than 0.442 but closer than 0.476 to the reference barcode. On the other hand, when we compute the confidence intervals using the reference barcode the empty set, we find that both the 95% and 99% confidence intervals are the cutoff value of 2. We find similar results for hypothesis barcodes with fewer than five bars.

We interpret these results to suggest that the hypothesis barcode is the best summarization amongst barcode distributions that put all of their mass on a single barcode. Of course, these results also suggest that when sampling at 500 points, we simply do not expect to see a distribution that is heavily concentrated around a single barcode. In the next subsection, we discuss the use of the witness complex, which does result in such a distribution.

*Remark 8.1.* To validate the non-parametric estimate of the confidence interval, we also used bootstrap resampling to compute bootstrap confidence intervals. Although we do not justify or discuss further this procedure herein (see instead [1]),



we note that we observed the reassuring phenomenon that the bootstrap confidence intervals agreed closely with the non-parametric estimates for both the 95% confidence intervals and the 99% confidence intervals in each instance.

**8.3. Results with the witness complex.** Because of the size of the datasets involved, in the analysis performed in [3], rather than the Rips complex VR they used the weak witness complex. The weak witness complex for a metric space  $(X, \partial)$  depends on a subset  $W \subset X$  of witnesses; the size of the complexes is controlled by  $|W|$  and not  $|X|$ .

**Definition 8.2.** For  $\epsilon \in \mathbb{R}$ ,  $\epsilon \geq 0$  and witness set  $W \subset X$ , the weak witness complex  $W_\epsilon(X, W)$  is the simplicial complex with vertex set  $W$  such that  $[v_0, v_1, \dots, v_n]$  is an  $n$ -simplex when for each pair  $v_i, v_j$ , there exists a point  $p \in X$  (a witness) such that the distances  $\partial(v_i, p) \leq \epsilon$ .

When working with the witness complex, we adapt our basic approach to study the induced distribution on barcodes which comes from fixing the point cloud and repeatedly sampling a fixed number of witnesses. The theoretical guarantees we obtained for the Rips complex in this paper do not apply directly; we intend to study the robustness and asymptotic behavior of this process in future work. Here, we report preliminary numerical results.

Again, we use as the hypothesis barcode the multi-set  $\{(0, c), (0, c), (0, c), (0, c), (0, c)\}$  as above. We approximated  $\text{MHD}_1^n(\mathcal{M}(15, 30))$  (various  $n$ ) by the empirical distribution obtained by sampling  $n$  landmark points from  $\mathcal{M}(15, 30)$ , computing the barcode, and computing the distance to the reference barcode. Doing this 1000 times, we find using the non-parametric estimate based on the interval statistics that the 95% confidence interval for  $\text{MHD}_1^{100}(\mathcal{M}(15, 30))$ , is  $[0.024, 0.027]$ . The 99% confidence interval for  $\text{MHD}_1^{100}(\mathcal{M}(15, 30))$  was also  $[0.024, 0.027]$ . When we computed the 95% confidence interval for  $\text{MHD}_1^{150}(\mathcal{M}(15, 30))$  we obtained  $[0.021, 0.023]$ . The 99% confidence interval for  $\text{MHD}_1^{150}(\mathcal{M}(15, 30))$  was  $[0.021, 0.024]$ . This represents high confidence for the data to be further than 0.021 (for  $\Phi_1^{150}$ ) and 0.024 (for  $\Phi_1^{100}$ ) but closer than 0.024 (for  $\Phi_1^{150}$ ) and 0.027 (for  $\Phi_1^{100}$ ) to the reference barcode. We obtained essentially the same results  $\text{MHD}_1^{500}$  as for  $\text{MHD}_1^{150}$ . We interpret these results to mean that the underlying distribution is essentially concentrated around the hypothesis barcode; the distance of 0.025 is essentially a consequence of noise.

*Remark 8.3.* In contrast, when we compute  $\text{MHD}_1^{25}(\mathcal{M}(15, 30))$ , we find the confidence interval is  $[1.931, 1.939]$ . When we compute  $\text{MHD}_1^{75}(\mathcal{M}(15, 30))$ , we find that the confidence interval is  $[1.859, 1.866]$ . This represents high confidence that  $\text{MHD}_1^{25}$  and  $\text{MHD}_1^{75}$  are far from this reference barcode, which in light of the confidence intervals above for  $\text{MHD}_1^{150}$  and  $\text{MHD}_1^{500}$  appear to indicate that samples sizes 25 and 75 are too small.

## REFERENCES

- [1] A. J. Blumberg, M. A. Mandell, Resampling methods for estimating persistent homology. In preparation.
- [2] A. J. Blumberg, I. Gal, and M. A. Mandell, Topological summarization of distributions of barcodes. In preparation.
- [3] G. Carlsson, T. Ishkhanov, V. de Silva., A. Zomorodian. On the local behavior of spaces of natural images. *International journal of computer vision*, 76 (1):1–12, 2008.
- [4] G. Carlsson and F. Memoli. Characterization, stability, and convergence of hierarchical clustering methods. *Journal of machine learning research*, 11:1425–1470, 2009.

- [5] G. Carlsson and V. De Silva. Zigzag persistence. *Foundations of computational mathematics*, 2009.
- [6] F. Chazal, D. Cohen-Steiner, L.J. Guibas, F. Memoli, S. Oudot. Gromov-Hausdorff stable signatures for shapes using persistence. *Comput. Graph. Forum*, 28(5):1393–1403, 2009.
- [7] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer. Stability of persistence diagrams. *Disc. and Comp. Geom.*, 37(1):103–120, 2007.
- [8] W.J. Conover. Practical nonparametric statistics. Wiley, 3rd edition, 1999.
- [9] H. A. David and H. N. Nagaraja. Order statistics. Wiley, 3rd edition, 2003.
- [10] H. Edelsbrunner, D. Letscher, and A. Zomorodian. Topological persistence and simplification. *Disc. and Comp. Geom.*, 28:511–533, 2002.
- [11] A. Greven, P. Pfaffelhuber, and A. Winter. Convergence in distribution of random metric measure spaces. *Prob. Theo. Rel. Fields*, 145(1):285–322, 2009.
- [12] J.A. Hartigan. Consistency of single linkage for high-density clusters. *J. Amer. Statist. Assoc.*, 76(374):388–394, 1981.
- [13] J.A. Hartigan. Statistical theory in clustering. *J. Classification*, 2(1):63–76, 1985.
- [14] M. Kahle. Random geometric complexes. *Disc. and Comp. Geometry*, 45(3):553–573, 2011.
- [15] M. Kahle and E. Meckes. Limit theorems for Betti numbers of random simplicial complexes. Preprint, arXiv:1009.4130, 2010.
- [16] J. Latschev. Vietoris-Rips complexes of metric spaces near a closed Riemannian manifold. *Archiv der Math.*, 77(6):522–528, 2001.
- [17] F. Memoli. Gromov-Wasserstein distances and the metric approach to object matching. *Foundations of Computational Mathematics*, 2011.
- [18] Y. Mileyko, S. Mukherjee, and John Harer. Probability measures on the space of persistence diagrams. *Inverse Problems*, 2012.
- [19] P. Niyogi, S. Smale, and S. Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Disc. and Comp. Geometry*, 39:419–441, 2008.
- [20] D. Pollard. Convergence of stochastic processes. Springer-Verlag, 1984.
- [21] V. de Silva and G. Carlsson. Topological estimation using witness complexes. *Proc. of Symp. on Point-Based Graph.*, 2004.
- [22] K-T. Sturm. On the geometry of metric measure spaces. *Acta Mathematica*, 196(1):65–131, 2006.
- [23] A. Zomorodian and G. Carlsson. Computing persistent homology. *Proc. of 20th ACM Symp. Comp. Geo.*, 2004.

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF TEXAS AT AUSTIN, AUSTIN, TX 78712

*E-mail address:* blumberg@math.utexas.edu

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF TEXAS AT AUSTIN, AUSTIN, TX 78712

*E-mail address:* igal@math.utexas.edu

DEPARTMENT OF MATHEMATICS, INDIANA UNIVERSITY BLOOMINGTON, IN 47405

*E-mail address:* mmandell@indiana.edu

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF TEXAS AT AUSTIN, AUSTIN, TX 78712

*E-mail address:* mpancia@math.utexas.edu